DOCUMENT RESUME

ED 446 071 SP 039 508

AUTHOR Baume, David; Yorke, Mantz
TITLE Validity and Reliability in the Evaluation of Portfolios for the Accreditation of Teachers in Higher Education.
PUB DATE 2000-06-17
NOTE 13p.; Paper presented at the American Association for Higher Education Assessment Forum (June 17, 2000). Page 11 is cropped.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Faculty; Evaluation Methods; Foreign Countries; Higher Education; *Portfolio Assessment; *Reliability; *Teacher Certification; *Teacher Evaluation; *Validity
IDENTIFIERS Authentic Assessment; United Kingdom

ABSTRACT

This paper examines the use of portfolio evaluation in accrediting higher education teachers in the United Kingdom. It explains portfolios as an approach to authentic assessment and discusses the issue of reliability in portfolio assessment when teacher accreditation depends on it. Researchers investigated one course, collecting data on 53 assessments from college records. The portfolios of 20 of those were regraded by trained assessors. The original 53 assessments were subjected to various statistical tests. The percentages of exact and close agreement were computed for six groups of items from the 74 comprising portfolio assessment. The experimental assessors marked more harshly than the original assessors, so subsequent findings were restricted to the original assessments. Interrater correlations ranged from -0.010 to 0.67. Patterns of agreement between the pairs of raters were significantly different from what would have been expected as a result of chance. Data showed problems with the components oriented toward analysis of needs and planning for future professional development. The valuing of equal opportunities was shown to be problematic. Assessors' judgments on the topic may vary considerably. (Contains 16 references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 446 071

Validity and reliability in the evaluation of portfolios for the accreditation of teachers in higher education.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. Yorke

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

David Baume, Centre for Higher Education Practice, The Open University, UK
(a.d.baume@open.ac.uk)

Mantz Yorke, Centre for Higher Education Development, Liverpool John Moores University, UK
(m.yorke@livjm.ac.uk)

[Note: this paper was presented as an introduction to a workshop session whose purpose included the discussion of the findings: hence no discussion is provided. The authors have written up this work more fully for formal publication.]

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Introduction

An approach to authentic assessment

Portfolios are now widely used for a variety of purposes relating to teacher development and appraisal, including self-presentation for promotion or tenure, personal accreditation as a teacher (Seldin, 1997), and accountability to super-ordinate authorities (such as school systems). Portfolios are also used for development and assessment in other professions including social work and nursing (Taylor et al, 1999). Portfolios are seen as 'authentic', in that they refer to collections of performances in naturalistic settings[1]. For that reason they are held to have advantages over other forms of assessment. However, Herman et al (1993, p.202) observed that 'the measurement quality of portfolios is largely uncharted territory': although there was a brief flurry of work on the reliability of portfolio assessments in the mid-1990s, it seems to have petered out without having advanced assessment methodology to any great extent.

Portfolios in the accreditation of HE teachers in the UK

The Open University in the UK runs courses for teachers in higher education, courses which lead to their accreditation as teachers and to a post-graduate qualification. The courses are accredited by the

[1] Though there is variation in expectation as to what should be included in a portfolio (Stecher, 1998; Simon and Forgette-Giroux, 2000).

SP039508

BEST COPY AVAILABLE

UK Staff and Educational Development Association (SEDA, 2000). SEDA specifies the seven or eight (depending on the course) outcomes to be achieved and the six principles or values which must demonstrably underpin the achievement of these outcomes. These are listed in Appendix 1. The teacher is required to present a portfolio of evidence in support of his/her claim for accreditation. The portfolio contains two distinct types of material; *evidence* (lesson plans, graded student work and the like), and *claims* in which the course participant argues that s/he has met the outcomes. A portfolio for the course which forms the object of this study requires a total of 74 assessment judgements. Twenty of these are technicalities such as word and page count. Forty-six require academic judgement on whether particular elements are demonstrated in the portfolios. The overall judgement on each of seven outcomes is obtained in major part by combining judgement on the elements of that outcome, with limited discretion left to the assessor and scope for the marginal failure of one element in an outcome to be condoned. In order to be judged to have passed the course, the final judgement, the teacher has to achieve a pass on all 7 outcomes.

This large number of elements of assessment results in part from the two-dimensional matrix of 7 outcomes, the attainment of each of which may need to be underpinned by up to 6 principles or values. The number is greater than 42 because each of the 7 outcomes is subdivided into components. For example, the needs to reflect on one's practice, identify one's development needs and plan one's continuing professional development are all subsumed under Outcome 7.

The Institute of Learning and Teaching (ILT, 2000), which is becoming the national organisation for the accreditation of teachers in higher education in the UK, envisages an assessment scheme which would, in effect, comprise a three-dimensional matrix of five outcomes or areas of work (such as teaching) by six areas of knowledge and understanding (such as models of how students learn) by five professional values (such as a commitment to equality of educational opportunity), giving potentially double the number of elements of assessment compared with that for OU Course H851. It is unlikely that such complexity of assessment will in practice be required.

Reliability is important

The potential for enormous growth in the use of portfolios by teachers in higher education makes both the validity and the reliability of portfolio assessment critically important topics. Reliability is of obvious importance when a teacher's accreditation depends on the assessment of his or her portfolio. In this summary, the primary focus is upon the issue of reliability. It is acknowledged that the validity of this assessment needs further attention, since it is determined, inter alia, by the curriculum design, the sampling of evidence, and the assessment methodology.

Reliability

Setting aside the issue of the selection of material for the portfolio (which – as noted above – is a validity issue), problems stem in the main from the judgement process that takes place. If judges do not agree about a portfolio's merits, then the assessment is unreliable. In her review of the use of portfolios in the UK's system of National Vocational Qualifications, Wolf (1998) indicates that a number of commentators have raised doubts about the reliability of portfolio assessment. Her conclusion, which is consistent with other findings relating to the problematic of shared understanding amongst teachers and assessors, is that it is impossible to develop written descriptions that are so tight that they can be applied reliably by multiple assessors to multiple assessment situations (Wolf, 1998, p441). There is (though it is admittedly difficult to identify) an optimal degree of precision in the specification of portfolio assessment tasks – too precise, and the detail makes the fulfilment of the tasks and the assessment unworkable in practice; too vague, and the whole process lacks focus.

Some technical comments on reliability

In this summary, exact agreement is differentiated from 'close agreements' which exist when the raters differ by no more than one point on the grading scale being used. Close agreement has been used in a number of the studies reported below.

Exact agreement is a stringent criterion, but (like the weaker 'close agreement') has to be interpreted against the likelihood that it could have been a result of chance. It is easy to work out the distribution of (dis)agreements that would be observed if each pair of possible judgements had an equal chance of appearing, and hence to test whether the observed distribution of (dis)agreements is significantly different from chance. Simply, the proportion of pairs of assessment judgements which would agree if assessment judgements were random is 1/N where N is the number of (equal-sized) bands on the grading scale used.

The interrater correlation coefficient is vulnerable to systematic differences between markers (which may raise it) and to minor variation in a sequence of broadly similar judgements (which lowers it).

Findings from studies of portfolio assessment

Reliability studies of the assessment of schoolchildren's portfolios and of portfolios put together by college students indicate respectable levels of interrater agreement [Appendix 2], although differences in presentation of the data from different studies make comparisons difficult.

These interrater agreements have been typically achieved in circumstances in which there was a template of defined outcomes against which to judge. Judgements become difficult to make when there is insufficient information: for example, 'outsider' judges of children's performances found greater difficulty than did the children's teachers in making judgements, presumably because the teachers had a fuller knowledge of their pupils and could hence interpolate the missing data (Supovitz et al, 1997). Further, bias is possible when the assessor picks up from the portfolio cues about the assessee (Howell et al, 1993).

The studies of reliability suggest that, whilst it may be possible to secure a reasonable level of interrater agreement in the assessment of portfolios, the underlying 'scatter' of gradings (evidenced in the correlational data) could be tightened up. Koretz (1998) suggests, however, that raters may not be the

largest sources of unreliability, and points to the sampling of tasks as being an important variable in this respect[2].

It can be reasonably concluded that reliability is enhanced when there are explicit outcome standards against which to judge (but Wolf's, 1998, stricture indicates that there is a limit to the degree of precision which can be encapsulated in the process of specification), and also when there are clear and unambiguous performance data upon which to exercise that judgement.

Nystrand et al's (1993) work indicates that reliability may be higher when assessors grade a set of portfolios by taking one element at a time and grading all students on that element before moving to the next element, than when they work their way through one portfolio before turning to the next. The 'element by element' approach may not always be practicable, and hence all that is likely to be possible is to maximise the reliability of assessments made *seriatim* through individual portfolios.

Assessing portfolios from OU HE Teacher Accreditation Courses

Background

Strenuous efforts are made to ensure valid and reliable assessment on these courses. The intended learning outcomes and underpinning principles and values are made explicit. Detailed guidance is given to participants and assessors on the meaning of these outcomes and underpinnings, and on the assessment process and standards. Examples are provided to participants of claims and evidence, together with assessors' comments. The assessors get together before each assessment to mark sample portfolios and develop shared understanding and further guidance. Through these processes, the aim is to maximise intersubjective agreement on what is required for assessment. The intersubjectivity extends to participants on the taught version of these courses, who receive tutor feedback on early drafts of sections of their portfolios.

[2] Note that, in such cases, reliability may be being compromised by a lack of definition in expectation regarding portfolio content, which has validity connotations.

Although – and as would be expected from the way the courses are run – the pass rate of students is high, the assessment process revealed through double-marking of portfolios that assessors did not always agree on their ratings of work[3]. A third assessor is brought in to resolve the major difficulties (particularly those at the pass/fail boundary). A study of interrater reliability was therefore undertaken in order to pinpoint where problems tended to be found, and hence to inform enhancement activities.

Method

The Course H851 was selected for investigation. Detailed data relating to 53 assessments were obtained from OU records. As an experiment, the portfolios of 20 of these were extracted from the archive and each was re-graded by two trained assessors who had not seen them previously. These assessors were asked to comment on the criteria being brought to bear when they were undertaking their assessments.

The original 53 assessments were subjected to various statistical tests.

The percentages of exact and close agreement were computed for six groups of items from the 74 making up the portfolio assessment:

- A and B: technicalities such as word and page counts
- Ci: gradings of the elements of the assessment such as components within outcomes
- Cii: evidence about the employment of the underpinning values
- D: the seven overall outcomes
- E: the course as a whole.

The interrater correlations (Pearson r) were computed for the graded components Ci and Cii.

The distributions of interrater (dis)agreements were computed in respect of components Ci and Cii, and compared with the distribution based on an 'equal chance' expectation, using the Kolmogorov-Smirnov one-sample test (Conover, 1971).

---

[3] Technicalities are graded pass/fail, but other aspects of the portfolios are more finely graded. The 46 components of outcomes are graded on a 4-point scale (well achieved; just achieved; not quite achieved; not achieved), and the 7 outcomes themselves are graded on a 5-point scale (outstanding pass; clear pass; bare pass; bare fail; clear fail).

Findings

The 20 'experimental assessments'

The 'experimental assessors' were found to have tended to mark more harshly than the original assessors of the 20 portfolios. This may be connected with the request that they comment on the assessment as they worked through the portfolio. This raises questions about the assessment method that cannot be answered by the evidence: does grading with overt reference to criteria perturb the normal approach and, if so, which of the two approaches is the more valid?

Because of this discrepancy in the experimental data, subsequent findings are restricted to the original assessments which are 'uncontaminated' by a possible 'experiment effect'.

The 53 original assessments

The pattern of exact and close agreements, by category, are shown in Table 1.

	A&B Claim size, etc	Ci Outcome elements	Cii Underpinning values	D Overall outcomes	E Overall result
Percentage exact agreement	96	64	55	39	60
Percentage close agreement	96	93	93	88	60

Table 1 Percentages of exact and close agreement for 53 portfolios.
The results for columns A, B and E are identical as these refer only to pass/fail judgements.

Since the underpinning values intersect all of the seven outcomes, it was possible to identify the pattern of discrepancies in assessing them relative to the pass/fail boundary, which is the critical one for course members (Table 2).

	OUTCOME							Mean
Underpinning value	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
How students learn	2	3	3	7			5	4.0
Concern for student development	3	2	4	4			7	4.0
Scholarship				8			8	8.0
Equal opportunities	8	11	9	9			8	9.0
Colleagueship				2		8	6	5.3
Reflection		7	6	4	7	6	5	5.8
Mean	**4.3**	**5.8**	**5.5**	**5.7**	**7.0**	**7.0**	**6.5**	**5.9**

Table 2. Discrepancy rates for underpinning values from original assessments of 53 portfolios, set against the seven outcomes. Differences counted in this table are those exceeding one grade.

The interrater correlations ranged from –0.10 to 0.67, with a median of 0.24.

The values of Dmax for the Kolgomorov-Smirnov one-sample tests ranged from 0.22 to 0.44 with a median of 0.36 (all significant at the .01 level [except one significant at the .05 level], one-tailed test on the grounds of *a priori* expectation). This set of tests shows that the patterns of agreement between the pairs of raters are significantly different from what would have been expected as a result of chance.

The data point towards problems with Outcome 7, and in particular with the components oriented towards analysis of needs and planning for future professional development. The valuing of equal opportunities is also shown to be problematic, perhaps because course members are faced with very different situations as far as equal opportunities are concerned, leading to variation in judgements about an acceptable level of performance. Assessors' judgements on the topic may also vary considerably.

References

Conover W J (1971) **Practical nonparametric statistics**. New York: Wiley.

Herman J L et al (1993) Assessing writing portfolios: issues in the validity and meaning of scores. **Educational Assessment 1**, pp201-224.

ILT (2000) The Insitute for learning and teaching in Higher Education. York:ILT. Available at http://www.ilt.ac.uk

Howell K W et al (1993) Bias in authentic assessment. **Diagnostique 19**, pp.387-400.

Koretz D et al (1993) **Interim Report: The reliability of Vermont portfolio scores in the 1992-93 school year**. Santa Monica, CA: RAND Institute on Education and Training.

Koretz D (1998) Large-scale portfolio assessments in the US: evidence pertaining to the quality of measurement. **Assessment in Education 5** (3), pp309-334.

LeMahieu P et al (1995) Portfolios in large-scale assessment: difficult but not impossible. **Educational Measurement: Issues and Practices 14**, pp.11-16 and 25-28.

Nystrand M et al (1993) Addressing reliability problems in the portfolio assessment of college writing. **Educational Assessment 1**, pp.53-70.

SEDA (2000) The SEDA Teacher Accreditation Scheme. Birmingham: SEDA. Available at http://www.seda.demon.co.uk/taccred.html

Seldin, P (1997) **The Teaching Portfolio**. Bolton, Mass: Anker Publishing.

Simon M and Forgette-Girouz R (2000) Impact of a content selection framework on portfolio assessment at the classroom level. **Assessment in Education 7** (1), pp83-101.

Stecher B (1998) The local benefits and burdens of large-scale portfolio assessment. **Assessment in Education 5** (3), pp335-351.

Supovitz J A et al (1997) Assessing agreement: an examination of the interrater reliability of portfolio assessment in Rochester, New York. **Educational Assessment 4**, pp. 237-259.

Taylor I et al (1999) Portfolios for learning and assessment: laying the foundations for continuing professional development. **Social Work Education 18** (2), pp147-160.

Wolf A (1998) Portfolio assessment as national policy: the national Council for Vocational Qualifications and its quest for a pedagogical revolution. **Assessment in Education 5** (3), pp413-445.

Wolfe E W (1996) A report on the reliability of a large-scale portfolio assessment for language arts, mathematics and science. ERIC ED 399 285.

Acknowledgement

We gratefully acknowledge the help that Martin Coffey gave us regarding data preparation.

Appendix 1 - Outcomes and underpinning principles and values for the course described

The **course outcomes** are that participants should be able to...

1 Plan teaching sessions

Design teaching sessions from a course outline, document or syllabus. This involves choosing teaching methods appropriate to the group of learners, the mode of study, the subject material, the resources available and the learning outcomes

2 Teach

Use two appropriate teaching and learning methods, and use appropriate learning technologies.

3 Assess student work

Mark or grade, and give feedback on, student work.

4 Monitor and evaluate their teaching

Monitor and evaluate your own teaching, using self, peer and student feedback.

5 Keep records

Keep appropriate records of your teaching support and academic administrative work.

6 Cope

Develop personal and professional coping strategies appropriate to the constraints and opportunities of your institutional setting, to manage adequately your time and operate successfully within available resources.

7 Continue your professional development

Reflect on your own personal and professional practice and development, assess your future development needs, and make a plan for your continuing professional development.

...and do so underpinned by these **principles and values**

1 How students learn

All teaching and academic administration should be informed by an understanding of how students learn and the conditions and processes that support student learning.

2 Concern for students' development

Helping students to learn must begin with a recognition that all students have their own individual learning needs and bring their own knowledge and resources to the learning process. Work with students should empower them and enable them to develop greater capability and competence in their personal and professional lives.

3 Commitment to scholarship

At the base of professional teaching is an awareness and acknowledgement of the ideas and theories of others. All teaching should be underpinned by a searching out of new knowledge - both about the subject/discipline and about good teaching and learning practice. All teaching should also lead to students developing a questioning and analytical approach.

4 Commitment to work with and learn from colleagues

Much of an academic's work is carried out as part of a team made up of teaching staff and academic support staff. The colleagueship and support of peers is as important as individual academic excellence.

5 Practising equal opportunities

Teachers must be concerned that students have equal opportunities, irrespective of disabilities, religion, sexual orientation, race or gender. So, everything that teachers do should be informed by equal opportunities legislation, by institutional policy and by a knowledge of best practice.

6 Continuing reflection on practice

Teachers should reflect on their intentions and actions and on the effects of their actions. They try to understand the reasons for what they see and for the effects of their actions. They thus continue to develop their understanding and practice and therefore inform their own learning.

Appendix 2 - Findings from a number of US studies of portfolio assessment

Study	Reliability measure	Comment
Herman et al (1993)	Interrater agreement ranged from 89% to 100% between pairs from 3 raters. Pearson r values ranged from 0.41 to 0.94.	Ratings cover whole portfolios and also components. 1 grade difference taken as agreement.
Koretz et al (1993)	Spearman rho correlations between raters around 0.60 for overall portfolio ratings.	Ratings for *components* of portfolios were lower.
Koretz (1998)	Initial interrater correlations 0.76 to 0.89, with near-perfect close agreement. 2 years later, 0.59 to 0.68.	Reporting on the National Assessment of Educational Progress Portfolio Assessment Trials, sampling Grade 4 and Grade 8 pupils.
Nystrand et al (1993)	Interrater agreement (Ns vary from 7 to 109) on portfolio elements ranged from 19% to 71%. Pearson r values ranged from –0.35 to 0.66.	Portfolios assessed sequentially; 1 grade difference taken as agreement.
	Interrater agreement (Ns vary from 48 to 493) on portfolio elements ranged from 53% to 79%. Pearson r values ranged from 0.44 to 0.86.	Items in portfolios assessed across all assessees' elements rather than portfolios assessed as wholes. Same criterion of agreement.
LeMahieu et al (1995)	Interrater correlations ranged between 0.74 and 0.87.[4]	Middle school and secondary level writing portfolios.
Heller et al (1998)	Percentage of exact agreement ranged from 48 to 63, and of close agreement from 91 to 100. Interrater reliability coefficients ranged from 0.53 to 0.83.	Ns ranged from 5 to 13; total N=84. Involved holistic ratings of portfolios from Grade 4 and Grade 8 pupils.
Supovitz et al (1997)	Spearman rho correlations between classroom teachers and external raters ranged between 0.58 and 0.77 (reading) and between 0.68 and 0.73 (writing). Corresponding ranges of N were 80-103 and 108-137.	Portfolio assessments were for kindergarten to Grade 2 classes.
Wolfe (1996)	Interrater correlations –0.04 to 0.55; 0.47 to 0.79, and 0.46 to 0.96 for science, language arts, and mathematics work samples, respectively. Respective exact agreement ranges were 33-64; 34-61; and 43-91: close agreements were 87-98; 80-93 and 80-100.	Secondary school pupils' work.

[4] Koretz (1998) argues that these figures are inflated due to inappropriate use of the Spearman-Brown prophecy formula in calculating reliability. According to Koretz, the real figures are 0.60-0.67 and 0.71-0.77 for high school and middle school portfolios respectively.